DOCUMENT RESUME

ED 449 678 FL 026 580

AUTHOR Nakamura, Yuji
TITLE Many-Facet Rasch Based Analysis of Communicative Language Testing Results.
PUB DATE 2000-03-00
NOTE 13p.; Published by Tokyo Keizai University, Association for Communication Studies.
PUB TYPE Journal Articles (080) -- Reports - Research (143)
LANGUAGE English, Japanese
JOURNAL CIT Journal of Communication Students; v12 p3-13 Mar 2000
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; *Communicative Competence (Languages); *English (Second Language); *Factor Analysis; Foreign Countries; Higher Education; *Language Tests; *Oral Language; Second Language Instruction; Second Language Learning; Tables (Data)
IDENTIFIERS Japan; Rasch Scaled Scores

ABSTRACT

This highly technical, data-rich paper examines how the Many-Facet Rasch Measurement Model can be applied to communicative language test data analyses and how beneficial the model is to language teachers. The results for a 20-item conversational response test and a 15-item sociolinguistic test of oral communication ability in English for 30 university students in Japan are presented and evaluated. (Contains 15 references.) (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 449 678

Many-Facet Rasch Based Analysis of Communicative Language Testing Results

By

Yuji NAKAMURA

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Prof. Yuji Nakamura

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

コミュニケーション科学第12号所載抜刷（2000年3月）

東 京 経 済 大 学

FL246580

Many-Facet Rasch Based Analysis of Communicative Language Testing Results

Yuji Nakamura

This paper examines how the Many-Facet Rasch Measurement Model can be applied to communicative language test data analyses and how beneficial the model is to language teachers. The data set used for this analysis is as follows (see Appendix for more details):

Test: Communicative Language Test
Items: 20 items (Conversational Response Test) and 15 items (Sociolinguistic Test)
Test Takers: 30 college students
Raters:6 English teachers at college
Rating Scale: (1:no answer, 2: inappropriate, 3:appropriate, 4:very good)

Linacre (1994) introduces the readers to the Rasch Measurement Model for judging as follows:

The construction of linear measures from qualitative observations is a conceptual and statistical advance of recent vintage. Rasch (1960/1980) obtains it by examinees for dichotomous responses to test items. Andrich (1978) and Masters (1982) expand the Rasch model to responses in ordered categories, e.g. attitude surveys and partial credit test items. This previous work has focused on observations resulting from the interaction of two components or "facets", objects and agents. In practice, a third facet is often encountered: a judge, rater or grader whose task is to award a rating to an examinee based on performance on a test item. The "many-facet Rasch model" extends the Rasch model to situations in which more than two facets interact to produce an observation. It enables the construction of a frame of reference in which quantitative comparisons no longer depend on which examinee happened to be rated by which judge on what item. (p.1).

Since the present paper deals with the assessment of students' oral communication ability or speaking ability, the involvement of raters (in other words, at least the third facet) cannot be avoided. Therefore, the many-facet Rasch measurement model can be of

great use for the language teachers in the following analysis. This present paper will examine the following five things:

1) the results of a many-facet Rasch analysis of the speaking test data (students, raters, items of the test on one common linear scale)
2) the student measurement report
3) the item measurement report
4) the rater measurement report
5) the category measurement report

First, we will look at Table 1. Table 1 shows that the students, raters and items of the oral language test data have been measured on one common linear scale. The expected scores (in rating points in the right hand column) are shown for students facing items of 0 logit difficulty and judges of 0 logit severity. Other expected scores are obtained by indexing the score scale (student ability--item difficulty--rater severity) at logits (cf. Linacre 1994). When we compare the student column and the item column, we see that for the top (able) student, all the items are easy to handle, and that even for the least able student, a few items are easy to manage. Thus, Table 1 provides us with quick and appropriate information of the relationship among students, items and raters on the common continuum.

Table 2 demonstrates the student measurement report. First, let us look at the column of Model Measure where we can investigate each student's ability by the Rasch measured logit scores more closely. This column shows that Student 4 (with the logit score of 2.17) is the most able student while Student 2 (with the logit score of -1.54) turns out to be the least able person. Next, we look at the mean square infit and outfit statistics, and find out that student 15, is a misfitting person against the expected value of 1 (cf. the acceptable range for mean square statistics is between .7 and 1.3). This student is worth examining. Finally, we look at the student separation index of 9.91 and reliability of .99, which is rather high. Since this index is equivalent to KR-20, Cronbach Alpha and the Generalizability coefficient, it can be said that the internal consistency reliability of the test for these students is reasonable.

Table 1 All Facet Vertical "Rulers".

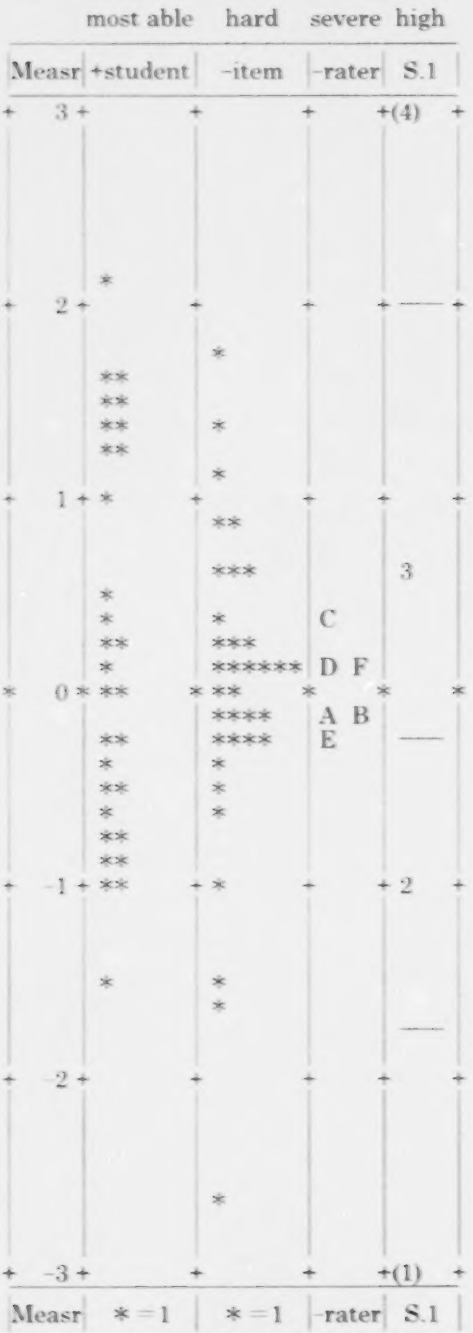

Table 2 Student Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	student
739	210	3.5	3.56	2.17	.13	1.2	1	1.3	2	4	12
701	210	3.3	3.38	1.61	.12	1.3	2	1.4	3	8	16
699	210	3.3	3.37	1.59	.12	1.1	0	1.2	1	17	24
696	210	3.3	3.35	1.55	.11	1.0	0	1.0	0	1	1
693	210	3.3	3.34	1.51	.11	1.2	1	1.1	1	26	5
683	210	3.3	3.29	1.38	.11	1.3	2	1.3	2	30	9
679	209	3.2	3.29	1.37	.11	1.0	0	1.0	0	10	18
677	210	3.2	3.26	1.31	.11	1.0	0	1.0	0	27	6
665	209	3.2	3.23	1.21	.11	0.8	-2	0.8	-2	3	11
654	210	3.1	3.16	1.04	.11	1.3	2	1.2	1	16	23
599	210	2.9	2.91	.47	.10	1.0	0	0.9	0	22	29
586	210	2.8	2.85	.35	.10	1.1	1	1.2	1	23	3
578	210	2.8	2.81	.28	.09	0.8	-2	0.8	-2	5	13
574	210	2.7	2.79	.25	.09	1.0	0	1.0	0	24	30
556	210	2.6	2.70	.09	.09	0.9	-1	0.9	-1	21	28
542	209	2.6	2.65	-.01	.09	1.0	0	1.0	0	12	2
542	210	2.6	2.63	-.03	.09	0.7	-4	0.7	-4	13	20
517	210	2.5	2.51	-.23	.09	0.8	-2	0.8	-2	9	17
511	209	2.4	2.49	-.26	.09	1.1	1	1.1	1	6	14
503	210	2.4	2.45	-.34	.09	0.7	-3	0.7	-3	14	21
483	210	2.3	2.33	-.50	.09	1.0	0	1.0	0	18	25
481	210	2.3	2.32	-.51	.09	0.9	-1	0.9	-1	19	26
463	209	2.2	2.23	-.63	.09	1.0	0	1.0	0	29	8
458	210	2.2	2.19	-.69	.09	0.8	-2	0.9	-1	7	15
449	209	2.1	2.15	-.74	.09	1.4	4	1.4	3	15	22
432	207	2.1	2.09	-.83	.09	0.8	-2	0.8	-2	11	19
440	210	2.1	2.09	-.83	.09	1.0	0	0.9	-1	28	7
421	209	2.0	1.99	-.97	.09	0.9	-1	0.8	-1	25	4
418	210	2.0	1.97	-1.00	.09	0.9	-1	0.8	-2	20	27
351	208	1.7	1.63	-1.54	.10	1.3	2	1.3	2	2	10
Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	student
559.7	209.6	2.7	2.70	.24	.10	1.0	-0.2	1.0	-0.2	Mean (Count: 30)	
106.4	0.7	0.5	0.52	.99	.01	0.2	2.1	0.2	2.0	S.D.	

RMSE (Model) .10 Adj S.D. .98 Separation 9.91 Reliability .99
Fixed (all same) chi-square: 2681.2 d.f.: 29 significance: .00
Random (normal) chi-square: 29.0 d.f.: 28 significance: .41

Table 3 shows another piece of useful information. First, let us look at the logit scores of Model Measure, and notice that item 14 (what is your) with the logit score of -2.57 is the easiest one, and that item 24 (will you) with the logit score of 1.80 is the most difficult one. Second, we examine the mean square infit and outfit statistics (cf. The expected value is 1 and the acceptable range is between .7 and 1.3), and find that all the items,

except items 24 (will you), 32 (how about), 15 (could you) and 23 (thank you), are within the acceptable range. These four items can be categorized as misfitting. Also, further investigation especially for items 24 and 15 is needed. Third, we need to consider the construct validity by looking at the order of difficulty of each item. The top six difficult items are from the second half of the conversation response test, while the bottom 5 easiest items are from the first half of the same test. In other words, these items in the conversation response test are well organized and presented in the order of difficulty, which is important from a psychological point of view.

Some other items from the conversation response test are mixed with the items from the sociolinguistic test.

The items from the Sociolinguistic test are assessing the intermediate level of ability of students, because most of the items from this test are located half way between the top difficult ones and the bottom easiest ones.

Furthermore, we need to consider whether or not the order of these items matches the theoretical construct of the oral proficiency.

Table 4 indicates the rater measurement report. One of the most salient parts of the Many-Facet Rasch analysis is that the Rasch approach recognizes, and models two aspects of judge (in the present research the term "rater" is used) behavior. Linacre (1994) claims that, first, judges are modelled to differ in severity or leniency. Second, he states that judges are modelled to exhibit, to some degree, stochastic behavior when awarding ratings. Linacre (1994) further confirms these two aspects by saying the following:

Even the most diligent judge training has failed to produce uniformity among judges (Borman,1978), but any difference among judges threatens fairness because the examinee raw score depends on which judge awards the rating. Indeed sometimes "there is as much variation among judges as to the value of each paper as there is variation among papers in the estimation of each judge" (Ruggles, 1911). It was this lack of judge reliability that was identified as a chief drawback to judge-dependent tests (Ruch, 1929). Clearly, since differences in judge severity can account for as much ratings variance as differences in examinee ability (Cason&Cason, 1984), objective measurement requires that judge behavior be modelled and statistically controlled. (p.6).

With these characteristics in mind, we look at Model Measure of the six raters. Rater E (-.28) turns out to be the most lenient while Rater C (.34) is the severest. The Infit and Outfit Mean Square Statistics indicates that all the raters fit the model. In other words, six raters consistently judge the students using the scale according to the Rasch model.

Table 3 Item Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	item
312	180	1.7	1.61	1.80	.11	1.6	4	1.3	1	24	will you
346	180	1.9	1.85	1.41	.10	1.0	0	0.9	0	28	I'd like
376	180	2.1	2.07	1.09	.10	0.9	0	0.9	-1	19	what is the
397	179	2.2	2.23	.87	.10	0.8	-2	0.8	-2	25	say hello
401	180	2.2	2.25	.84	.10	1.2	1	1.3	2	29	I'll
423	180	2.4	2.40	.62	.10	1.1	1	1.2	2	30	do you
426	180	2.4	2.42	.59	.10	0.8	-2	0.8	-2	36	questioning
429	180	2.4	2.44	.56	.10	0.8	-2	0.9	-1	47	information
445	180	2.5	2.55	.41	.10	0.9	0	0.9	0	45	opinion
459	180	2.5	2.63	.27	.10	1.2	2	1.2	1	27	let's have
460	179	2.6	2.65	.23	.10	1.1	0	1.1	1	39	disagreeing
464	180	2.6	2.66	.22	.10	1.4	3	1.3	3	32	how about
466	179	2.6	2.69	.18	.10	1.2	2	1.3	2	17	what do you
469	179	2.6	2.70	.15	.10	1.3	3	1.3	2	18	can you
471	180	2.6	2.70	.15	.10	0.9	0	0.9	0	42	warning
471	180	2.6	2.70	.15	.10	0.7	-3	0.8	-2	44	telephone
475	180	2.6	2.72	.11	.10	0.9	-1	0.9	-1	35	repetition
474	179	2.6	2.73	.09	.10	0.7	-4	0.7	-3	34	complaining
481	180	2.7	2.76	.05	.10	0.8	-1	0.8	-1	46	offering
483	180	2.7	2.77	.03	.10	0.6	-5	0.6	-4	41	interrupting
493	179	2.8	2.84	-.11	.10	0.9	-1	0.9	-1	37	greeting
498	180	2.8	2.85	-.13	.10	1.0	0	1.1	0	21	Sunday
494	178	2.8	2.86	-.14	.10	0.9	-1	1.0	0	40	congratulating
502	180	2.8	2.87	-.17	.10	0.9	-1	0.9	-1	22	how
504	180	2.8	2.88	-.19	.10	0.8	-2	0.9	-1	43	telephone
509	179	2.8	2.92	-.26	.10	1.4	3	1.4	3	15	could you
515	180	2.9	2.94	-.31	.10	1.1	0	1.1	0	31	would you
515	180	2.9	2.94	-.31	.10	0.9	0	0.9	0	38	parting
517	180	2.9	2.95	-.33	.10	0.6	-4	0.7	-3	33	apologizing
532	180	3.0	3.03	-.50	.11	1.1	1	1.0	0	26	it's a
540	180	3.0	3.07	-.60	.11	1.4	3	1.3	2	23	thank you
575	180	3.2	3.25	-1.04	.12	0.8	-2	0.8	-1	20	weather
601	179	3.4	3.41	-1.48	.13	-1.2	1	1.4	2	16	how are you
608	178	3.4	3.47	-1.65	.13	0.6	-3	0.7	-2	13	nice to meet
659	180	3.7	3.72	2.57	.16	0.9	0	1.2	1	14	what is your
Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Nu	item
479.7	179.7	2.7	2.73	.00	.11	1.0	-0.3	1.0	-0.2	Mean (Count: 35)	
69.0	0.6	0.4	0.41	.81	.01	0.2	2.5	0.2	2.2	S.D.	

RMSE (Model) .11 Adj S.D. .81 Separation 7.57 Reliability .98
Fixed (all same) chi-square: 1565.4 d.f.: 34 significance: .00
Random (normal) chi-square: 33.8 d.f.: 33 significance: .43

The separation index (4.75) is not as good as the ideal one (if there is a total agreement among raters, the separation index should be 0.00); however, this score can be acceptable in terms of the practical aspect of the performance test assessment.

Table 4 Rater Measurement Report (arranged by mN).

Obsvd Score	Obsvd Count	Obsvd Average	Fair Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	N	rater
2617	1048	2.5	2.59	.34	.04	1.0	-1	1.0	-1	3	C
2741	1050	2.6	2.72	.12	.04	1.0	0	1.0	0	4	D
2729	1046	2.6	2.72	.12	.04	0.8	-4	0.9	1	6	F
2868	1050	2.7	2.85	-.12	.04	1.0	0	1.0	0	2	B
2896	1050	2.8	2.87	-.17	.04	0.9	-3	0.9	-3	1	A
2939	1044	2.8	2.93	-.28	.04	1.3	6	1.3	5	5	E
2798.3	1048.0	2.7	2.78	.00	.04	1.0	-0.5	1.0	-0.1	Mean (Count: 6)	
111.9	2.3	0.1	0.12	.21	.00	0.1	3.4	0.1	2.8	S.D.	

RMSE (Model) .04 Adj S.D. .21 Separation 4.75 Reliability .96
Fixed (all same) chi-square: 141.3 d.f.: 5 significance: .00
Random (normal) chi-square: 5.0 d.f.: 4 significance: .29

Finally, Table 5 shows the category statistics. The four categories (1:no answer, 2:inappropriate, 3:appropriate, 4:very good) were well used to evaluate students performance, which can be explained in the columns of counts used and percentages.

Table 5 Category Statistics.

DATA Category Score	Counts Used	%	Cum. %	FIT Avge Meas	OUTFIT MnSq	STEP CALIBRATIONS Measure	S.E.	EXPECTATION Measure at Category	-0.5	MOST PROBABLE from	THURSTONE THRESHOLD at
1	1009	16%	16%	-1.18	1.0			(-2.46)		low	low
2	1287	20%	37%	-.39	.9	-1.03	.04	-.96	-1.78	-1.03	-1.45
3	2761	44%	80%	.48	1.1	-.77	.03	.68	-.24	-.77	-.43
4	1231	20%	100%	1.50	1.0	1.80	.04	(2.94)	1.98	1.80	1.86
								Mean		Modal	Median

In summary, we can draw some conclusions from the research:

1) The Many-Facet Rasch Analysis can give us many pieces of useful information, not only globally, but also locally.
2) The global information is given by the "birds' eye view" of three facets on the same continuum, where the relative position of 3 facets can be recognized quite easily.
3) The local information is provided by the item measurement report, the student

measurement report, the rater measurement report and the category statistics.

4) The rater measurement report, which is one unique part of the Many-Facet Rasch model, can lead researchers into a more detailed and thorough discussion of the performance test (involving raters) analysis.

Note: This research was supported in part by Tokyo Keizai University under Research Grant APC01-97.

Acknowledgement:

I am grateful to Dr. Benjamin D. Wright and Dr. John M. Linacre for their invaluable comments

Bibliography

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika* 43: 561-573.

Bode, R. K., and Wright, B. D. (1999). Rasch measurement in higher education. Reprinted from Higher Education: Handbook of Theory and Research, Vol. XIV (pp.287-316). New York: Agathon Press.

Borman, W. C. (1978). Exploring upper limits of reliability and validity in job performance ratings. *Journal of Applied Psychology* 63: 134-144.

Cason, G. J., and Cason, C. L. (1984). A deterministic theory of clinical performance rating. *Evaluation and the Health Professions* 7: 221-247.

Linacre, J. M. (1989, 1993, 1994). *Many-Facet: Rasch Measurement*. Chicago: MESA Press.

Linacre, J. M. (1998). *Facets Rasch Software Users Guide*. Chicago: MESA Press.

Linacre, J. M. and Wright, B. D. (1998). *Facets: Many-Faceted Rasch Analysis*. Chicago: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika* 47: 149-174.

Rasch, G. (1960, 1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, and Chicago: University of Chicago Press.

Ruch, G.M. (1929). *The objective or new-type examination*. Chicago: Scott, Foresman.

Ruggles, A. M. (1911). *Grades and grading*. New York: Teacher's College. Reported in F. J. Kelly, Teacher's marks. New York: Teacher's College. 1914.

Wright, B. D. (1997). A history of social science. *Educational Measurement: Issues and Practice*. Winter 1997, 33-45&52.

Wright, B. D. (1997). Fundamental measurement for psychology. In S. Emretson, and S. Hershberger (Eds.). *The New Roles of Measurement: What Every Psychologist and Educator Should Know* (pp.65-104). Hillsdale NJ: Lawrence Erlbaum Associates.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.

Wright, B. D, and Stone, M. H. (1979). *Best Test Design: Rasch Measurement. Chicago:* MESA Press.

Appendix A Conversational Response Test

Directions:

1) You will hear twenty questions or sentences in English each followed by a pause.
2) Please give a quick and appropriate response in English to each sentence.

Twenty record sentences or question:

13) Nice to meet you.
14) What is your name?
15) Could you spell it please?
16) How are you?
17) What do you do?
18) Can you tell me the time?
19) What is the date today?
20) Wha is the weather like today?
21) What do you usually do on Sunday?
22) How do you come to school?
23) Thank you for everything.
24) Will you do me a favor?
25) Say hello to your family.
26) It's beautiful day, isn't it?
27) Let's have a cup of coffee.
28) I'd like you to meet my sister.
29) I'll see you at the restaurant at six tomorrow.
30) Do you mind if I use your eraser?
31) Would you like some ice cream for dessert?

32) How about playing tennis next Sunday?

Appendix B Sociolingustic (Mini Contexts) Competence Test

Directions:

1) You will hear fifteen contexts in Japanese each followed by a pause.

2) Please give an appropriate response in English in each context.

33) Context 1 (Apologizing and making an excuse)

You are late for your class. You missed the school bus. Please apologize and make an excuse to your teacher.

34) Context 2 (Complaining and requesting)

You are in a non-smoking section of a waiting room at the airport. Someone started smoking. You have a cold and a sore throat. Please complain about it and request him/her to stop it.

35) Context 3 (Asking for repetition)

You didn't understand what your teacher said. You want the teacher to repeat it. Please make a request to your teacher.

36) Context 4 (Questioning)

You want to know the train schedule. Please ask about the departure time of the next train for Kyoto at the ticket office.

37) Context 5 (Greeting)

You happen to meet your high school teacher (Mr.Suzuki) after a long interval. Please greet him.

38) Context 6 (Parting)

After talking a while, you part from your teacher. Please say "farewell" to him.

39) Context 7 (Disagreeing)

Your friend (Tomoko) says jogging is a healthy activity. You doun't agree wiht her. What do you say to her?

40) Context 8 (Congratulation)

Your friend's older sister won the first prize in a speech contest. Please congratulate her on her success.

41) Context 9 (Interrupting)

Your supervisor is working in his office. You want to interrupt him for a

moment to talk with him. What do you say?

42) Context 10 (Warning)

Some children are playing baseball and almost break the window of your house. Please warn them.

43) Context 11 (Telephoning)

You are making a phone call. You want to speak to Mr. Brown. What do you say?

44) Context 12 (Telephoning)

You answer the phone. Someone wants to talk with your father. But he is out now. What do you say?

45) Context 13 (Getting an opinion)

You wnat to get your friend's opinion about last week's college festival. What do you say?

46) Context 14 (Offering)

You want to serve something to drink to a guest at your house. Please offer something to drink.

47) Context 15 (Asking for infomation)

At department store, please ask the receptionist where the stationery section is.

Appendix C Scoring Sheets for Two Tests

Conversational Response Test

	no answer	conversationally inappropriate	conversationally appropriate	very good
13
 32				

Sociolinguistic (Mini Contexts) Competence Test

	no answer	sociolinguistically inappropriate	sociolinguistically appropriate	very good
33
 47				